

RNA-Seq Demo on Galaxy

Tom Doak

Le-Shin Wu

Carrie Ganote

National Center for Genome Analysis Support

August 12, 2015



INDIANA UNIVERSITY



INDIANA UNIVERSITY

Our RNA-Seq Demo Data



Cristobal Rojas, La miseria (1886)

We will be assembling the DNA Polymerase protein units from the H37Rv strain of *Mycobacterium tuberculosis*, the causative agent of TB, also known as the consumption.

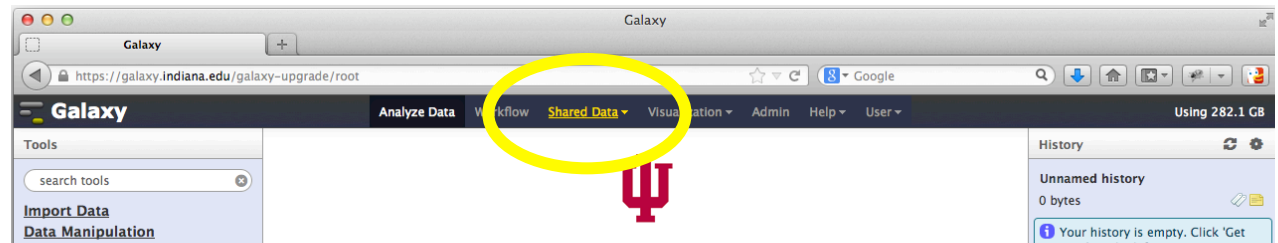
The raw reads originated from the Short Read Archive on NCBI. The accession number for the set is SRX212035.

This dataset consists of paired-end, ~75bp RNA-Seq reads.

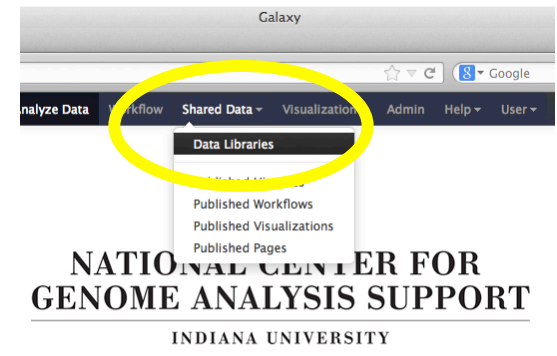


Let's get some sequence data

Galaxy allows users to publish their data to share with each other.



Let's start with "Shared Data" at the top.
Then select Data Libraries from the menu.





Let's get some sequence data

The screenshot shows the Galaxy web interface. At the top is a navigation bar with the Galaxy logo and links for 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', and 'Admin'. Below this is the 'Data Libraries' section, which includes a search bar with the placeholder text 'search dataset name, info, message, dbkey' and a magnifying glass icon. Below the search bar is a link for 'Advanced Search'. A table with two columns, 'Data library name' and 'Data library description', is displayed. The first row is 'User Import Library' with the description 'For moving large datasets into Galaxy'. The second row is 'Workshop Data' with the description 'Learning sets of RNA-Seq data'. The 'Workshop Data' link is circled in yellow.

<u>Data library name</u> ↓	<u>Data library description</u>
<u>User Import Library</u>	For moving large datasets into Galaxy
<u>Workshop Data</u>	Learning sets of RNA-Seq data

Choose Workshop Data.



Let's get some sequence data

Expand folder →

Check both boxes →

Name	Message	Data type
Galaxy Workshop September '13		
<input checked="" type="checkbox"/> TB_1.fq		fastqsanger
<input checked="" type="checkbox"/> TB_2.fq	Right reads	fastqsanger

For selected datasets: Import to current history Go

TIP: You can download individual library datasets by selecting "Download this dataset" from the context menu (triangle icon).

TIP: Several compression options are available when downloading multiple library datasets simultaneously:

- gzip: Recommended for fast network connections
- bzip2: Recommended for slower network connections (smaller size but takes longer to compress)
- zip: Not recommended but is provided as an option for those who cannot open the above formats

Import the Data sets to current history.



Let's get some sequence data

Galaxy Analyze Data Workflow Shared Data Visualizations

Data Library "Workshop Data"

✓ 2 datasets imported into 1 history: Unnamed history

<input type="checkbox"/> Name	Message
<input type="checkbox"/> Galaxy Workshop September '13	
<input type="checkbox"/> TB_1.fq	
<input type="checkbox"/> TB_2.fq	Right reads

For selected datasets: Import to current history Go

i TIP: You can download individual library datasets by selecting "Download this dataset"

Data set is imported – Click on Analyze Data to return.

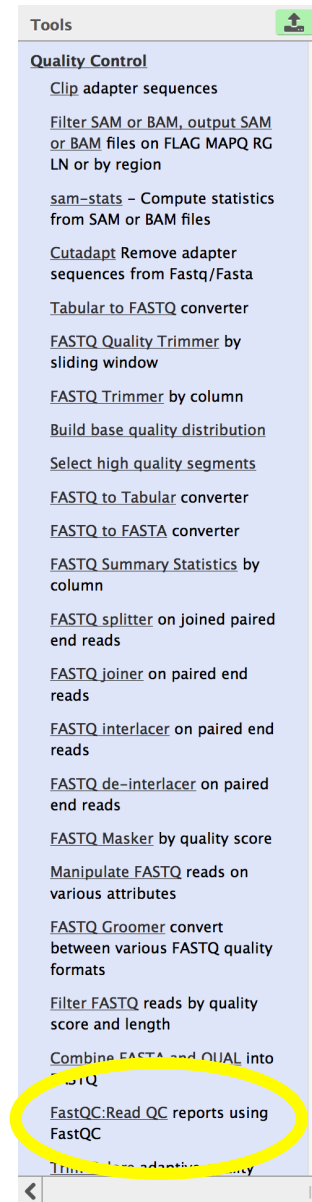


Step 1: Assess the Quality of Inputs

We will first get an idea of the quality of our input data sets.

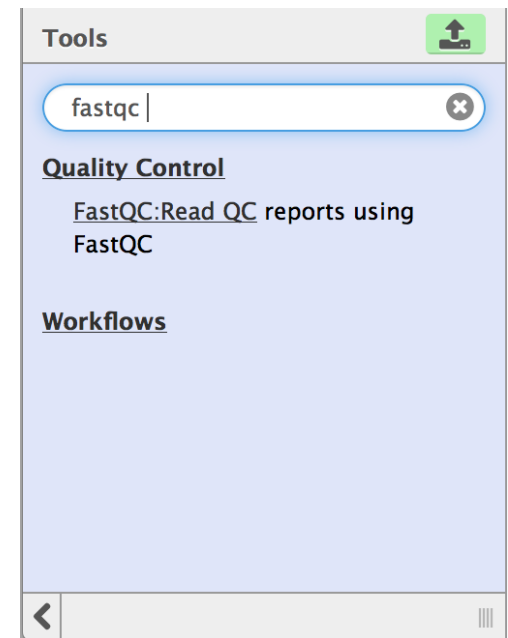
The FastQC tool will produce graphical output that makes it easy to gauge the characteristics of the data – quality, patterns, biases, gc content etc.

Choose the left or right reads file and run it. Compare your results with your neighbors’.



Pro tip: Use the search bar to find tools

(I added a space at the end here)



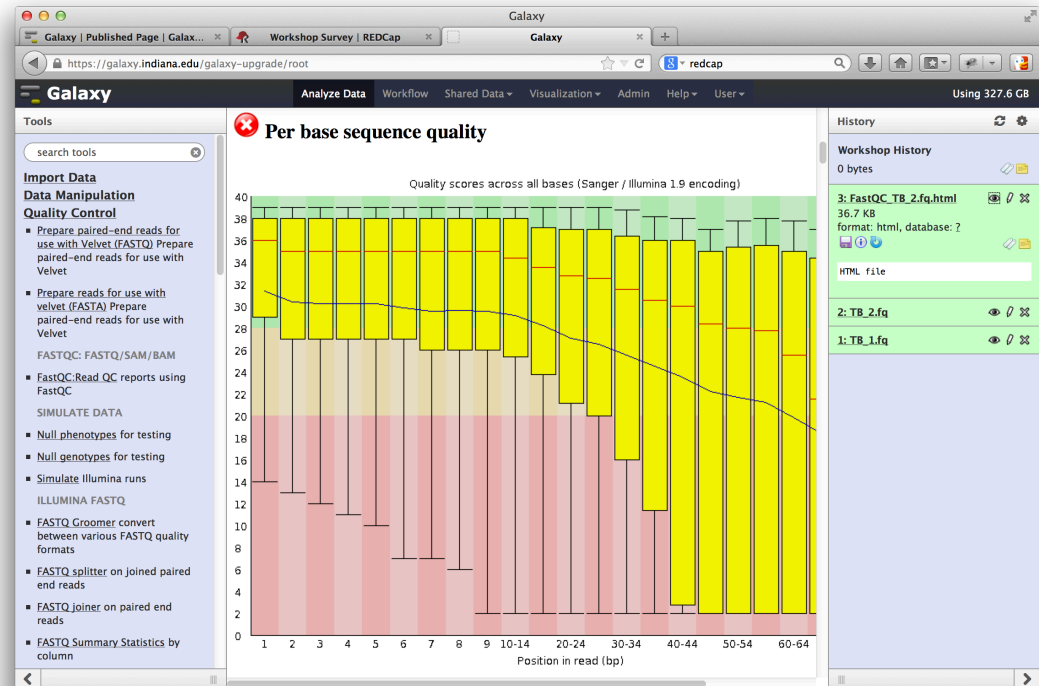


Step 1: Assess the Quality of Inputs

The input data usually declines in quality as the reads progress.

The quality score is assigned by the sequencing machine as it reads each base. It is a rough estimate of how ambiguous the signal is.

Sequence: **ATGCATG**
Quality Score: 39 38 23 19 3 3





Step 2: Trim Input Sequences

We've determined that the input data sets need some work before they are used in downstream processes. We'll use the FASTQ quality trimmer by sliding window to trim reads based on quality score.

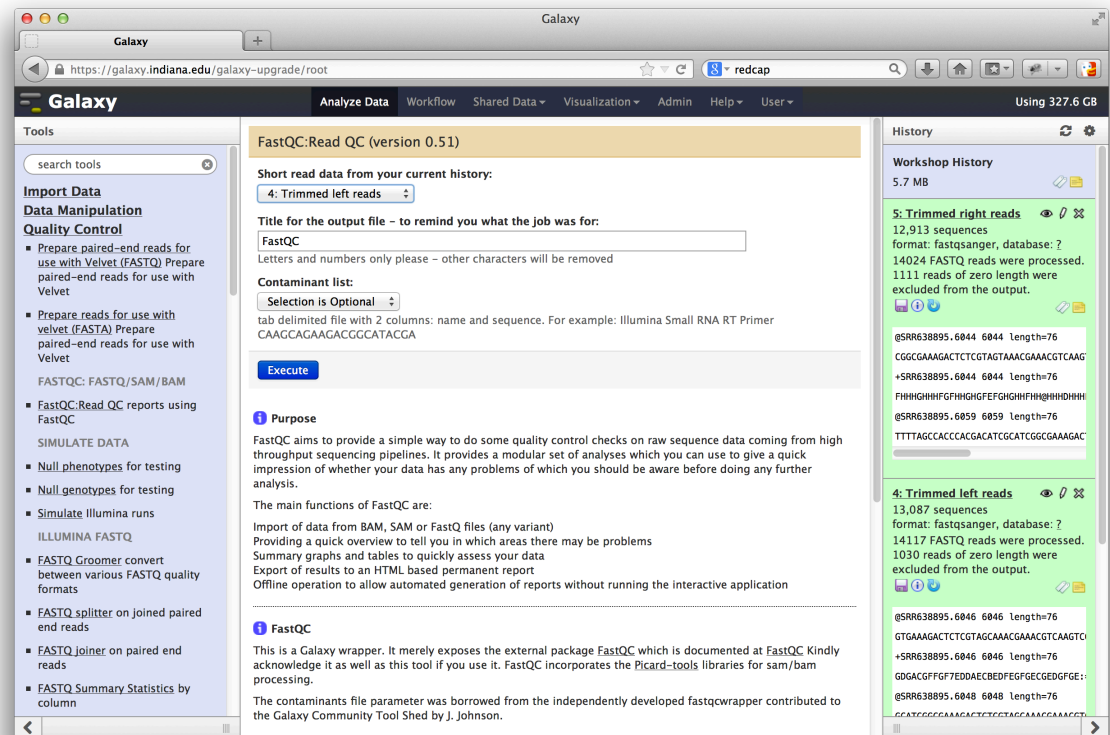
Run this tool for both input data sets.

The screenshot shows the Galaxy web interface. On the left, the 'Tools' panel lists various tools. The tool 'FASTQ Quality Trimmer by sliding window' is highlighted with a yellow circle. On the right, the tool's configuration page is shown. The title is 'FASTQ Quality Trimmer (version 1.0.0)'. The 'FASTQ File' is set to '1: TB_1.fq'. The 'Keep reads with zero length' checkbox is unchecked. The 'Trim ends' are set to '5' and 3''. The 'Window size' is set to '1'. The 'Step Size' is set to '1'. The 'Maximum number of bases to exclude from the window during aggregation' is set to '0'. The 'Aggregate action for window' is set to 'min score'. The 'Trim until aggregate score is:' is set to '>='. The 'Quality Score' is set to '20.0'. An 'Execute' button is at the bottom. Below the button, a description states: 'This tool allows you to trim the ends of reads based upon the aggregate value of quality score within a sliding window; a sliding window of size 1 is equivalent to 'simple' trimming of the ends of reads. The user specifies the aggregating action (min, max, sum, mean) to perform on the quality scores within the sliding window to be used with the user defined comparison operation and threshold.'



Step 3: Rinse, Repeat

Now that the files are trimmed, we will re-assess their quality. If necessary, keep trimming away until you are satisfied with the input files.



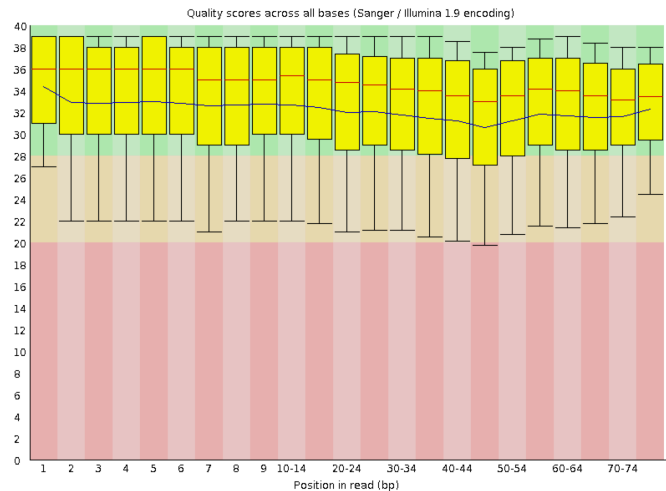
I renamed my trimmed files to help me keep them straight.



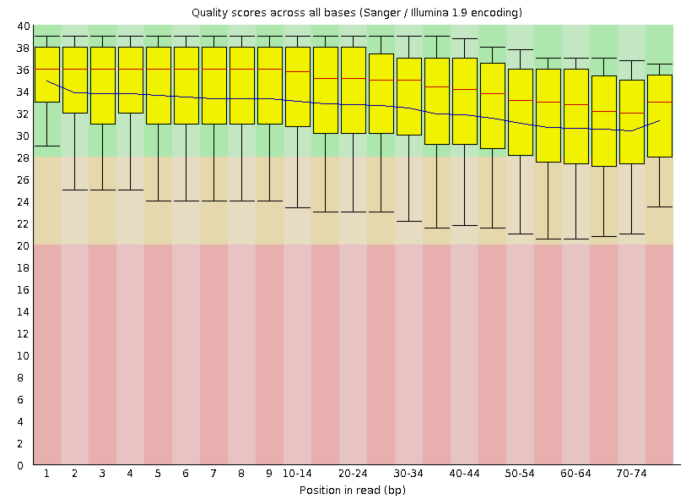
Step 3: Rinse, Repeat

Pictured are the left and right reads after trimming is complete.
These will do!

✓ Per base sequence quality



✓ Per base sequence quality





Step 4: Assembly

Next we will put the reads together to create a complete picture of the actively transcribed genes of the sample organism.

Trinity is a *de novo* assembler that has been optimized for use on Mason. We will use it to assemble our reads.

The screenshot displays the Galaxy web interface. On the left, a sidebar lists various tool categories: Tools, Import Data, Data Manipulation, Quality Control, De novo Assembly, Mapping and Alignments, Run Blast+, Run Blast+ on Open Science Grid, Annotation, Statistics, Variants, Clustering/Phylogeny, Visualization, and Workflows. The 'De novo Assembly' category is highlighted with a yellow oval, and the 'Trinity - Executes on Mason' tool is selected within it. The main panel shows the configuration for the Trinity tool (version 0.0.1). The configuration options include: 'Paired or Single-end data?' set to 'Paired'; 'Left/Forward strand reads' set to '4: Trimmed left reads'; 'Right/Reverse strand reads' set to '5: Trimmed right reads'; 'Strand-specific Library Type' set to 'None'; 'Paired Fragment Length' set to '300'; 'Is it strand specific data?' set to 'No'; 'Use Additional Params?' set to 'No'; and 'How long will your job need?' set to '1 hr'. An 'Execute' button is located at the bottom of the configuration panel.



It finished! We're done, right?

An assembler solves a computer problem of putting together a puzzle from tiny pieces. The output of the assembler is a guess – but we don't know how accurate it is. We could look at:

- Basic stats of the assembly – “Contigs”
 - Number of “Contigs” vs. Expected Number
 - N50 – a weighted average
 - Average Length
 - Max Length
- Check contigs against known genes with Blast

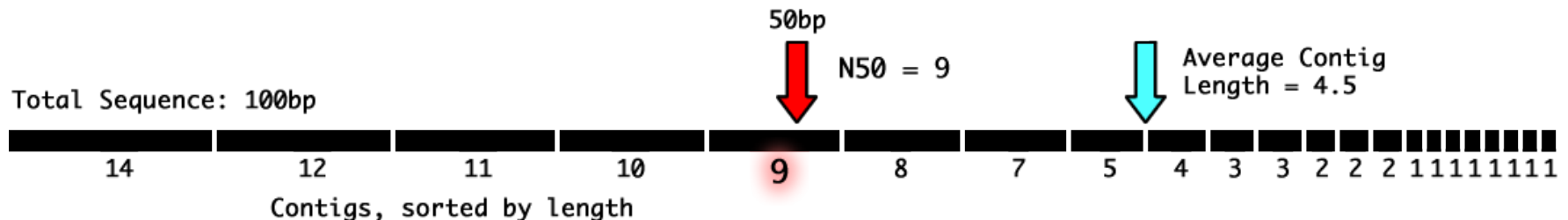
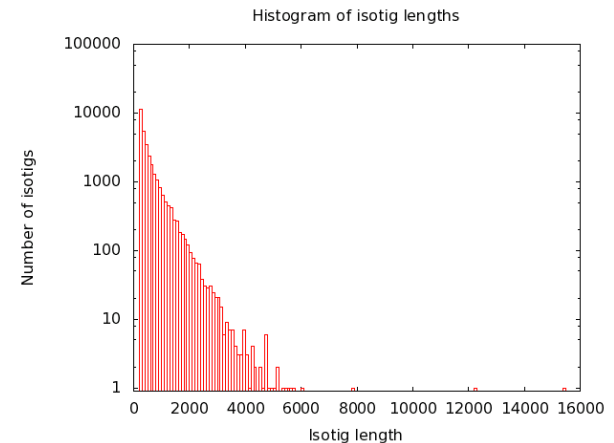


Step 5: Assessing Quality of Assembly

Important statistics for assembly quality:

Contig Length Distribution

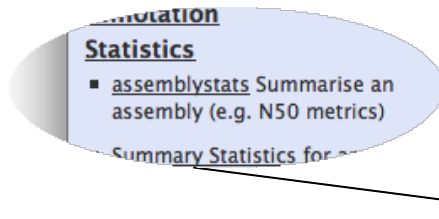
Assemblies will typically produce a number of complete contigs representing whole transcripts, and a large number of partial transcripts. This biases the average contig length toward the low end. The N50 is a measure weighted by total sequence length in the assembly.





Step 5: Assessing Quality of Assembly

Getting these stats in Galaxy:



Run assemblystats to get a summary and histograms of your contig length distribution.

The screenshot shows the Galaxy web interface with a workflow named 'assemblystats (version 1.0.1)'. The interface includes a top navigation bar with 'Galaxy' and various menu items. The left sidebar lists tools under 'Tools', with 'Statistics' highlighted. The main panel shows the 'assemblystats' tool configuration, including 'Type of read' (Isotig), 'Output histogram with bin sizes=1', and 'Source file in FASTA format' (84: Trinity on data 20 and data 21: Assembled Transcripts). The right sidebar shows the 'History' panel with two entries: '240: Sorted contigs' and '239: Assembly statistics'. The '239: Assembly statistics' entry is expanded, showing a table of statistics for isotig lengths.

1 2	
Statistics for isotig lengths:	
Min isotig length:	
Max isotig length:	
Mean isotig length:	
Standard deviation of isotig leng	
Median isotig length:	



Step 6: Getting more data

Right-click HERE and choose “Copy link location”.

Galaxy / at IU

Analyze Data Workflow Shared Data Visualization Admin Help User

Tools

search tools

Import Data

Data Manipulation

Quality Control

De novo Assembly

Mapping and Alignments

Run Blast+

Annotation

Statistics

Variants

Clustering/Phylogeny

Visualization

NGS: Mapping

Workflows

Download data directly from web or upload files from your disk

Name	Size	Type	Genome	Settings	Status
New File	0.1 KB	Auto-det...	unspecified (?)		100%

You can tell Galaxy to download data from web by entering URL in this box (one per line). You can also directly paste the contents of a file.

ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF_000195955.2_ASM19595v2/GCF_000195955.2_ASM19595v2_protein.faa.gz

You can Drag & Drop files into this box.

Choose local file Paste/Fetch data Start Pause Reset Close

The TB accession can be found at NCBI.



Step 7: Check Against Database

For this step, we'll check to see how well our assembled transcripts compare to what we already know.

Use this step to give a rough annotation of genes, to make sure that your transcripts are from nuclear genes, or to gauge how complete your sequence is.

Tools

- protein domain database (PSSMs) with protein query sequence(s)
- [NCBI BLAST+ database info](#) Show BLAST database information from blastdbcmd
- [NCBI BLAST+ blastp](#) Search protein database with protein query sequence(s)
- [NCBI BLAST+ rpstblastn](#) Search protein domain database (PSSMs) with translated nucleotide query sequence(s)
- [BLAST XML to tabular](#) Convert BLAST XML output to tabular
- [NCBI BLAST+ tblastx](#) Search translated nucleotide database with translated nucleotide query sequence(s)
- [NCBI BLAST+ dustmasker](#) masks low complexity regions
- [NCBI BLAST+ blastx](#) Search protein database with translated nucleotide query sequence(s)
- [NCBI BLAST+ makeblastdb](#) Make BLAST database**

Annotation

Statistics

Variants

NCBI BLAST+ makeblastdb (version 0.0.22)

Molecule type of input:

☒ protein

☐ nucleotide

FASTA files

FASTA file 1

file:

Add new FASTA file

Title for BLAST database:

This is the database name shown in BLAST search output

Parse the sequence identifiers:

☐

This is only advised if your FASTA file follows the NCBI naming conventions using pipe '|' symbols

Enable the creation of sequence hash values:

☒

These hash values can then be used to quickly determine if a given sequence data exists in this BLAST database.

Masking data files

Add new Masking data file

How long will your job need:

Execute



Step 7: Check Against Database

We will use Blastx to search the database for our genes.

Use blast database from history.

The screenshot shows the NCBI BLAST+ web interface. On the left, the 'Tools' menu is open, and 'NCBI BLAST+ blastx Search protein database with translated nucleotide query sequence(s)' is highlighted with a yellow circle. The main panel is titled 'NCBI BLAST+ blastx (version 0.0.22)'. It contains the following fields:

- Nucleotide query sequence(s):** 11: Trinity on data 7 and data 9: Assembled Transcripts
- Subject database/sequences:** BLAST database from your history
- Protein BLAST database:** 95: protein BLAST database from data 94
- Query genetic code:** 1. Standard
- Set expectation value cutoff:** 0.001
- Output format:** Pairwise HTML
- Advanced Options:** Hide Advanced Options
- How long will your job need:** 1 hr

An 'Execute' button is at the bottom of the configuration panel. Below the panel, there is an information icon and a note: 'Note. Database searches may take a substantial amount of time. For large input datasets it is advisable to allow overnight processing.'

Make sure to choose Pairwise HTML output for readability.



Step 7: Check Against Database

We see the expected genes as the top hits!

The screenshot displays the NCBI BLAST web interface. The top navigation bar includes links for Analyze Data, Workflow, Shared Data, Visualization, Admin, Help, and User. The main results pane shows a list of top hits with their accession numbers and descriptions. The first hit is DNA polymerase III subunit delta' [Actinomyces...]. Below the list, a detailed alignment is shown for the query sequence (WP_003907097.1) against the subject sequence (EFD75343.1). The alignment shows a high degree of identity (100%) and a score of 246 bits (627). The right-hand sidebar contains a History section with a list of previous searches, including '17: blastx on db' and '16: Filter sequences by length on data 9'.

Accession	Description
ref WP_009744290.1	DNA polymerase III subunit delta' [Actinomyces...]
ref WP_005963184.1	DNA-directed DNA polymerase III subunit delta' [Actinomyces...]
ref WP_010525702.1	DNA polymerase III subunit delta' [Nesterenkia...]
ref WP_017201448.1	hypothetical protein [Microbacterium barkeri]
ref YP_830214.1	DNA polymerase III subunit delta' [Arthrobacter...]
ref WP_019482356.1	hypothetical protein [Arthrobacter sp. TB 23]
ref WP_004806969.1	DNA polymerase III subunit delta' [Actinomyces...]

Query: WP_003907097.1 DNA polymerase III subunit delta, partial [Mycobacterium tuberculosis H37Rv]
 Subject: EFD75343.1 DNA polymerase III subunit delta [Mycobacterium tuberculosis H37Rv]
 Length=354

Score = 246 bits (627), Expect = 3e-77
 Identities = 181/181 (100%), Positives = 181/181 (100%), Gaps = 0/181
 Frame = -3

Query 553 TDPQARQRRERALGLARDAATPSRAYAAAEELVAGAEAEALALTAQRIEAEETEELRTA
 TDPQARQRRERALGLARDAATPSRAYAAAEELVAGAEAEALALTAQRIEAEETEELRTA
 Sbjct 174 TDPQARQRRERALGLARDAATPSRAYAAAEELVAGAEAEALALTAQRIEAEETEELRTA

Query 373 aggtgkgtgaalrgatgAMKDLERRQKSRQTRASRDALDRALIDLATYFRDALLVAAH
 AGGTGKGTGAALRGATGAMKDLERRQKSRQTRASRDALDRALIDLATYFRDALLVAAH
 Sbjct 234 AGGTGKGTGAALRGATGAMKDLERRQKSRQTRASRDALDRALIDLATYFRDALLVAAH

Query 193 GVRANHPDMADRVAALAAHAPPERLLRCIEAVLACREALAVNVKPKFAVDAMVATIGC
 GVRANHPDMADRVAALAAHAPPERLLRCIEAVLACREALAVNVKPKFAVDAMVATIGC
 Sbjct 294 GVRANHPDMADRVAALAAHAPPERLLRCIEAVLACREALAVNVKPKFAVDAMVATIGC

Query 13 R 11

We could limit the number of hits depending on output desired.



Step 8: Differential Expression

Tools
Annotation
Statistics
[Draw ROC plot on "Perform LDA" output](#)
[MINE - Maximal Information-based Nonparametric Exploration](#)
[Perform LDA Linear Discriminant Analysis](#)
[Generate A Matrix for using PC and LDA](#)
[Gene level DE test across two conditions](#) Runs EBSeq to find DE genes across two conditions
[Count GFF Features](#)
[Correlation for numeric columns](#)
RSEM prepare reference
[RSEM calculate expression RNA-Seq by Expectation-Maximization](#)
[RSEM trinity fasta to gene map extract transcript to gene map from trinity](#)

RSEM prepare reference (version 1.1.17)
Reference transcript source:
transcript fasta
reference fasta file: 11: Trinity on data 7 and data 9: Assembled Transcripts
The files should contain the sequences of transcripts.
Map of gene ids to transcript (isoform) ids: Selection is Optional
Each line of should be of the form: gene_id transcript_id (with the two fields separated by a tab character) The map can be obtained from the UCSC table browser group: Genes and Gene Prediction Tracks table: knownisoforms Without a map: If a reference genome and gtf is used, then RSEM uses the "gene_id" and "transcript_id" attributes in the GTF file. Otherwise, RSEM assumes that each sequence in the reference sequence files is a separate gene.
reference name:
rsem_ref_name
A one word name for this RSEM reference containing only letters, digits, and underscore characters
PolyA :
Do not add poly(A) tails to any transcripts
Disable the conversion of 'N' characters to 'G' characters in the reference sequences:
☐
Bowtie uses the automatic N to G conversion to to align against all positions in the reference.
Execute

We will use RSEM and EBSeq to calculate differential gene expression.

First we need to build a reference using RSEM prepare reference.



Step 8: Differential Expression

Next, gene counts will be produced using RSEM. Make sure to use the prepared library and the paired end reads. Do not create a BAM file.

Tools

- [Import Data](#)
- [Data Manipulation](#)
- [Quality Control](#)
- [De novo Assembly](#)
- [Mapping and Alignments](#)
- [Run Blast+](#)
- [Annotation](#)
- [Statistics](#)
 - [Draw ROC plot on "Perform LDA" output](#)
 - [MINE - Maximal Information-based Nonparametric Exploration](#)
 - [Perform LDA Linear Discriminant Analysis](#)
 - [Generate A Matrix for using PC and LDA](#)
 - [Gene level DE test across two conditions](#) Runs EBSeq to find DE genes across two conditions
 - [Count GFF Features](#)
 - [Correlation for numeric columns](#)
 - [RSEM calculate expression RNA-Seq by Expectation-Maximization](#)
 - [RSEM trinity fasta to gene map](#) extract transcript to gene map from trinity
 - [Get All Possible Patterns in a Multiple Condition Design](#) Get all possible patterns in a multiple condition design
 - [Get Normalized Expressions](#) Calculate normalization factors and get the normalized expression matrix
 - [Isoform level DE test across two conditions](#) Runs EBSeq to find

RSEM calculate expression (version 1.1.17)

Sample name:
rsem_sample

RSEM Reference Source:
From your history

RSEM reference:
20: RSEM rsem_ref_name reference

RSEM Input file type:
FASTQ

FASTQ type:
phred33 qualities (default for sanger)

Library type:
Paired End Reads

Read 1 fastq file:
1: TB_1.fq

Read 2 fastq file:
2: TB_2.fq

bowtie settings:
use bowtie defaults

Seed length used by the read aligner:
25
Providing the correct value for this parameter is important for RSEM's accuracy if the data are single-end reads. RSEM uses this value for Bowtie's seed length parameter. The minimum value is 25. (Default:25)

Is the library strand specific?:
No

Additional RSEM options:
Use RSEM Defaults

Create bam results files:
No BAM results files
In addition to the transcript-coordinate-based BAM file output, also output a BAM file with the read alignments in genomic coordinates

Execute



Step 8: Differential Expression

We will treat this sample like a real set with three replicates. The notation for the condition is tricky – C1 and C2 are sample names, and it assumes the data is in tabular with one gene name column and all other columns are counts.

Tools
Statistics
[Draw ROC plot on "Perform LDA" output](#)
[MINE – Maximal Information-based Nonparametric Exploration](#)
[Perform LDA Linear Discriminant Analysis](#)
[Generate A Matrix for using PC and LDA](#)
[Gene level DE test across two conditions](#) Runs EBSeq to find DE genes across two conditions
[Count GFF Features](#)

Gene level DE test across two conditions (version 1.0.0)
Gene Expression (tab delimited, please use the unnormalized values, e.g. expected counts form RSEM):
80: Compute on data 79
The First Row is Sample Names?:
No
Enter which condition each sample belongs to (separated by comma, no space please):
C1,C1,C1,C2,C2,C2,
Target FDR:
1
Execute



INDIANA UNIVERSITY

Step ..?

RNA-Seq is a very versatile technology. You can use the data for:

- Gene discovery based on transcripts
- Genome evidence – introns, exons, junction
- Gene expression patterns in multiple samples
- SNP calling/other variants
- Protein divergence between samples

We have gotten to the assembly step, but there is a lot to learn about the data now that it is put together. A foundation in the use of Galaxy coupled with Indiana University resources will enable you to reach these goals.



INDIANA UNIVERSITY

Fin

Thanks for watching!
Questions and comments:
Email help@ncgas.org